Survey Experiments

Business Experimentation and Causal Methods

Revealed vs Stated Preference

- Revealed preference:
  - What people actually do.
    - Buying products, donating, voting, etc…
- Stated preference:
  - What people say they would do.

Stated Preference Is Often Biased

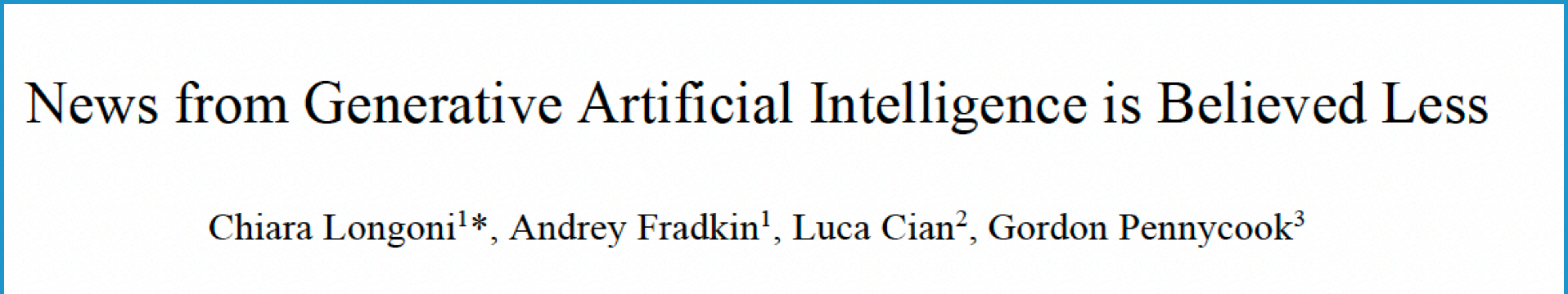
- Social desirability bias
- Hypothetical bias
- Strategic bias
- Information bias
- Preference uncertainty
- Lots of others

Why do survey experiments?

- Cost
- Speed
- More control over design
- Ethical considerations
- Access to subjective perceptions.
- What you’re interested may not have happened before.
    - 2024 election, self-driving cars, generative AI, etc…

Two main experiment types

- Across subjects
  - Randomization unit is the subject.
- Within subjects.
  - Randomization unit is subject by question.
  - Advantage is more power, disadvantage is spillovers.
    - Often called ‘carry-over effects’.



News from Generative Artificial Intelligence is Believed Less

Chiara Longoni[1]*, Andrey Fradkin[1], Luca Cian[2], Gordon Pennycook[3]

Research Question

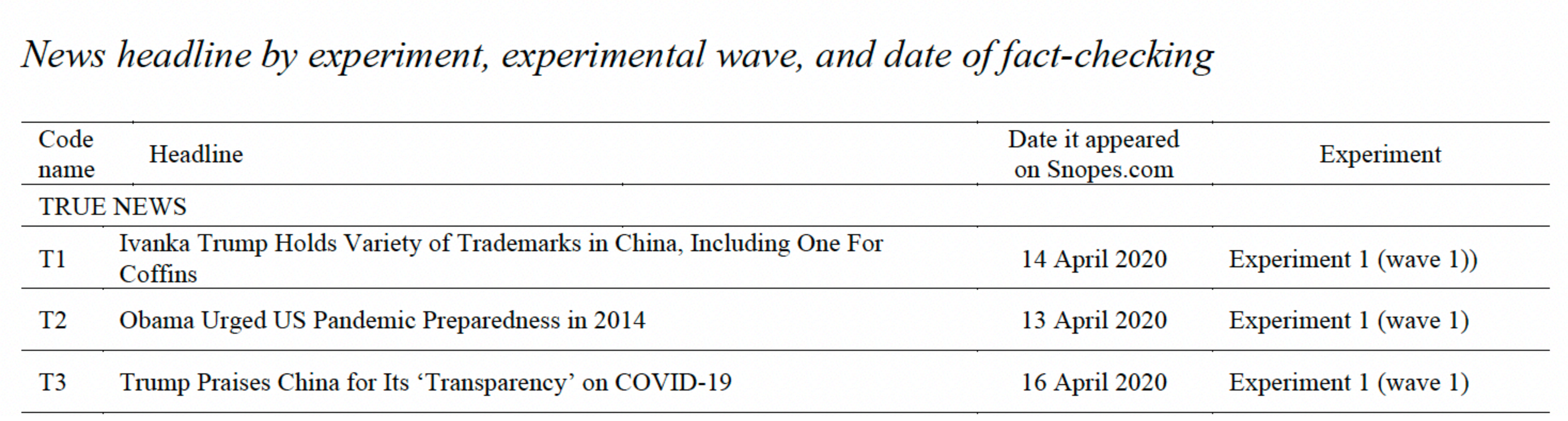
- Paper written prior to ChatGPT.
- What is the effect of labeling news articles as being written by AI on accuracy perceptions?
- Important question, since already some news articles were being generated.
- Various proposals to label such articles.

Setup

- Pick true and false headlines / photos from Snopes (site that tracks disputed articles). A lot of these were about COVID.

- Main Outcome:

  - Perception of accuracy (1 - Not at all accurate, 4 - very accurate).

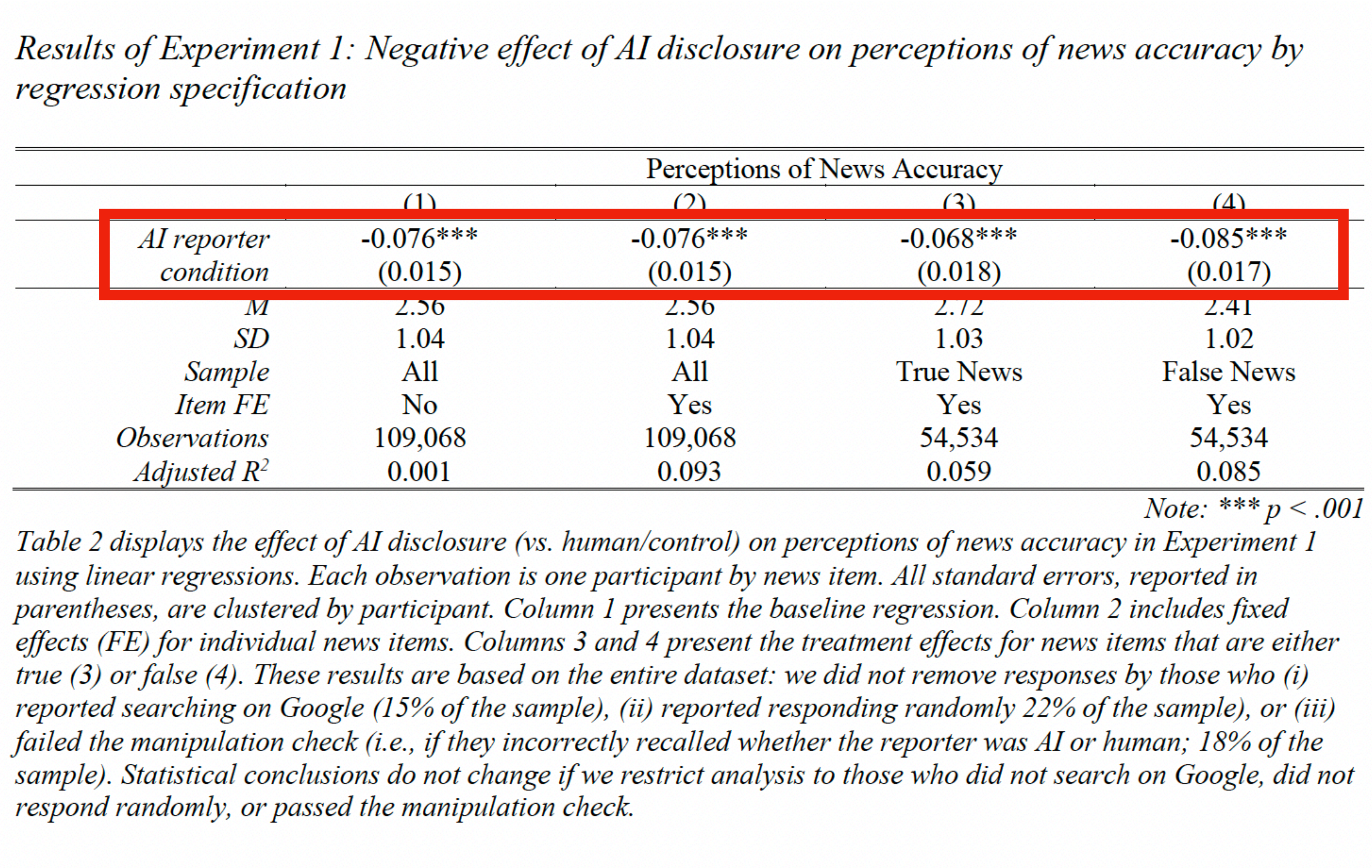


News headline by experiment, experimental wave, and date of fact-checking

Code name	Headline	Date it appeared on Snopes.com	Experiment
TRUE NEWS			
T1	Ivanka Trump Holds Variety of Trademarks in China, Including One For Coffins	14 April 2020	Experiment 1 (wave 1))
T2	Obama Urged US Pandemic Preparedness in 2014	13 April 2020	Experiment 1 (wave 1)
T3	Trump Praises China for Its 'Transparency' on COVID-19	16 April 2020	Experiment 1 (wave 1)

Across subject design.

- A 2-cell, between-subject design.
- 3,029 participants for Experiment 1.



Results of Experiment 1: Negative effect of AI disclosure on perceptions of news accuracy by regression specification

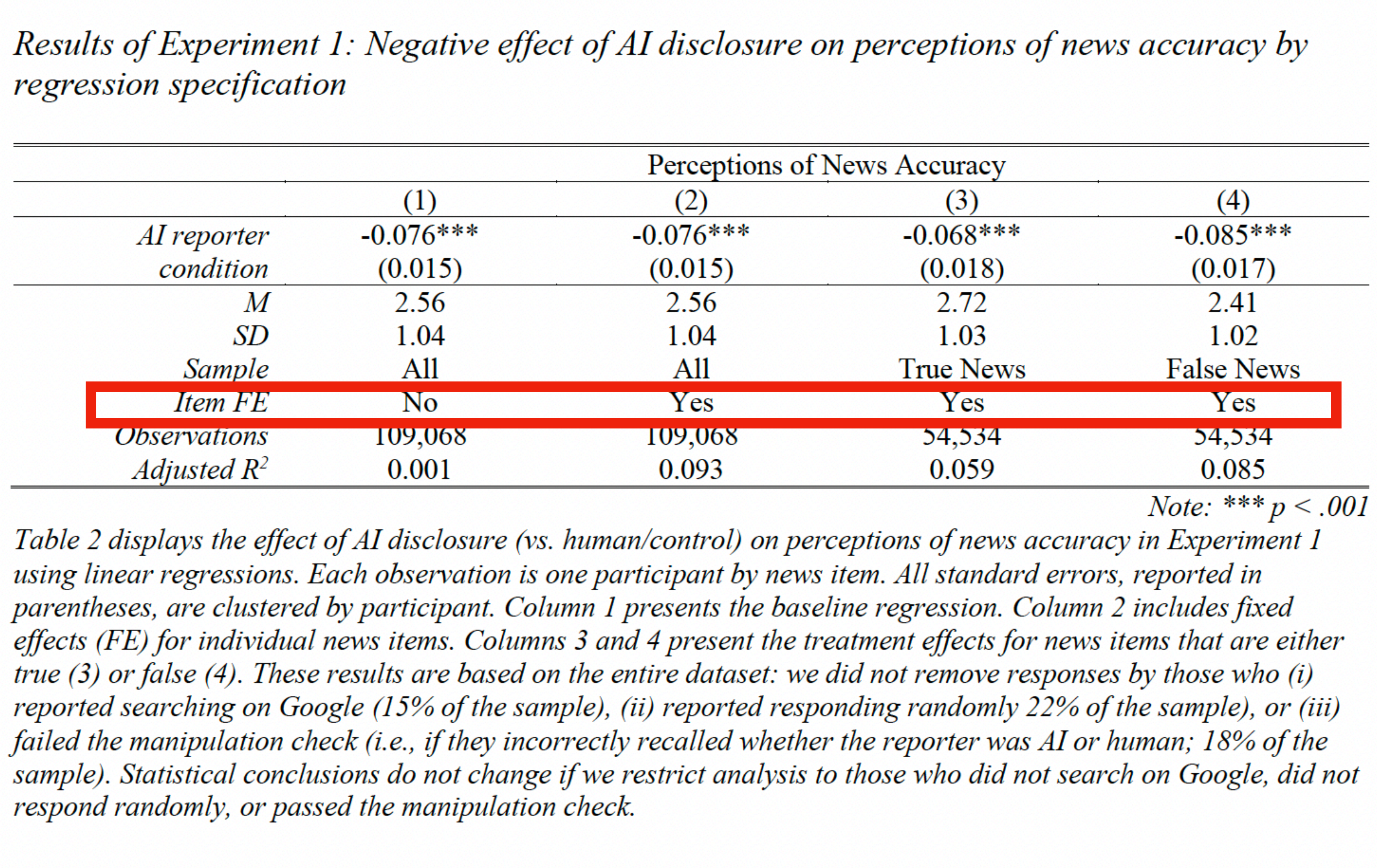
	Perceptions of News Accuracy			
	(1)	(2)	(3)	(4)
AI reporter condition	-0.076*** (0.015)	-0.076*** (0.015)	-0.068*** (0.018)	-0.085*** (0.017)
M	2.56	2.56	2.72	2.41
SD	1.04	1.04	1.03	1.02
Sample	All	All	True News	False News
Item FE	No	Yes	Yes	Yes
Observations	109,068	109,068	54,534	54,534
Adjusted R^2	0.001	0.093	0.059	0.085

*Note: *** $p < .001$*

Table 2 displays the effect of AI disclosure (vs. human/control) on perceptions of news accuracy in Experiment 1 using linear regressions. Each observation is one participant by news item. All standard errors, reported in parentheses, are clustered by participant. Column 1 presents the baseline regression. Column 2 includes fixed effects (FE) for individual news items. Columns 3 and 4 present the treatment effects for news items that are either true (3) or false (4). These results are based on the entire dataset: we did not remove responses by those who (i) reported searching on Google (15% of the sample), (ii) reported responding randomly 22% of the sample), or (iii) failed the manipulation check (i.e., if they incorrectly recalled whether the reporter was AI or human; 18% of the sample). Statistical conclusions do not change if we restrict analysis to those who did not search on Google, did not respond randomly, or passed the manipulation check.

Item Fixed Effects - Items are Headlines

- A 2-cell, between-subject design.
- 3,029 participants for Experiment 1.



Results of Experiment 1: Negative effect of AI disclosure on perceptions of news accuracy by regression specification

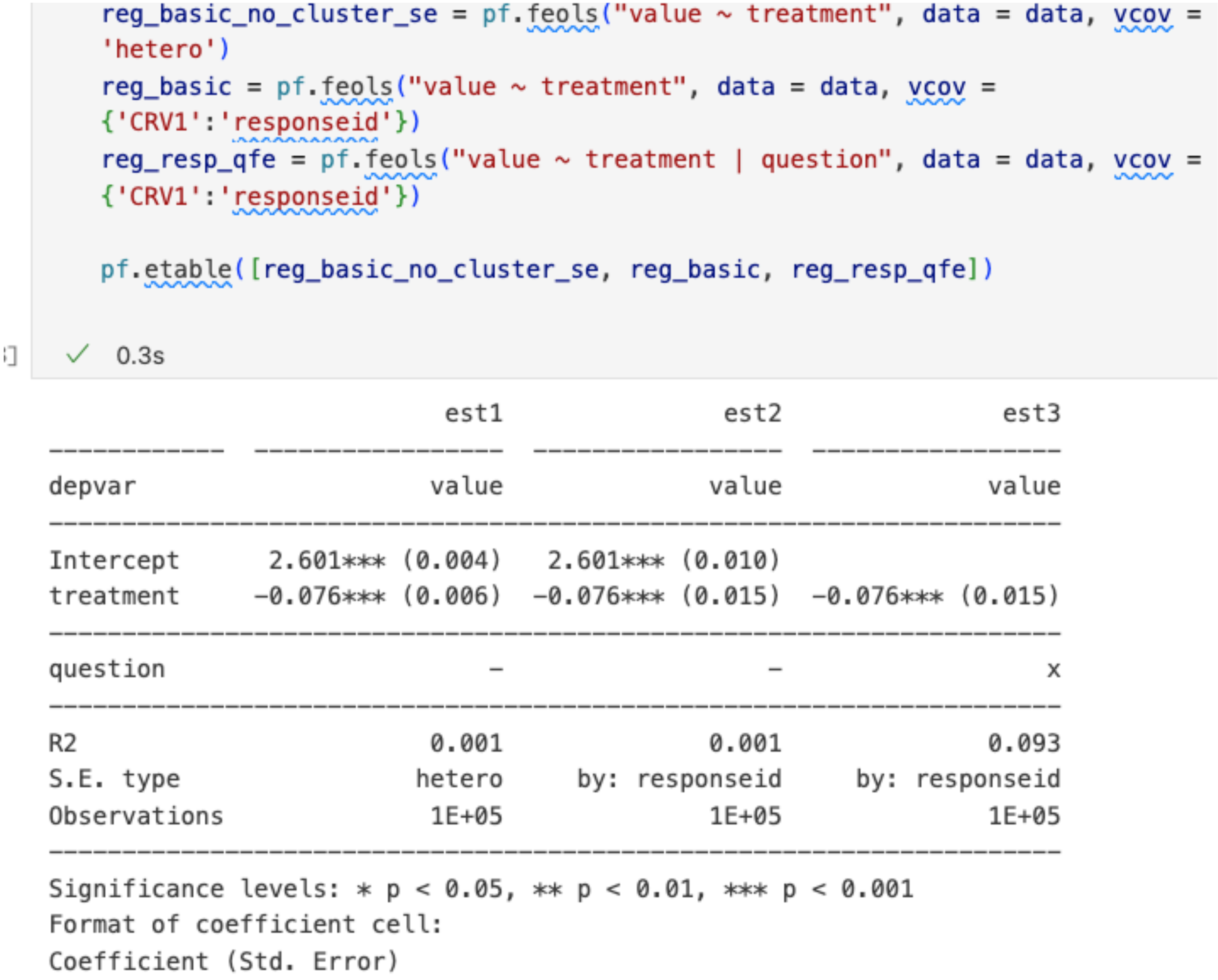
	Perceptions of News Accuracy			
	(1)	(2)	(3)	(4)
AI reporter condition	-0.076*** (0.015)	-0.076*** (0.015)	-0.068*** (0.018)	-0.085*** (0.017)
M	2.56	2.56	2.72	2.41
SD	1.04	1.04	1.03	1.02
Sample	All	All	True News	False News
Item FE	No	Yes	Yes	Yes
Observations	109,068	109,068	54,534	54,534
Adjusted R^2	0.001	0.093	0.059	0.085

*Note: *** $p < .001$*

Table 2 displays the effect of AI disclosure (vs. human/control) on perceptions of news accuracy in Experiment 1 using linear regressions. Each observation is one participant by news item. All standard errors, reported in parentheses, are clustered by participant. Column 1 presents the baseline regression. Column 2 includes fixed effects (FE) for individual news items. Columns 3 and 4 present the treatment effects for news items that are either true (3) or false (4). These results are based on the entire dataset: we did not remove responses by those who (i) reported searching on Google (15% of the sample), (ii) reported responding randomly 22% of the sample), or (iii) failed the manipulation check (i.e., if they incorrectly recalled whether the reporter was AI or human; 18% of the sample). Statistical conclusions do not change if we restrict analysis to those who did not search on Google, did not respond randomly, or passed the manipulation check.

Standard Errors: Cluster at unit of randomization or at higher level of aggregation.

- See ‘course_materials/survey_code/analyze_survey.ipynb’.
-



```
reg_basic_no_cluster_se = pf.feols("value ~ treatment", data = data, vcov =
'hetero')
reg_basic = pf.feols("value ~ treatment", data = data, vcov =
{'CRV1':'responseid'})
reg_resp_qfe = pf.feols("value ~ treatment | question", data = data, vcov =
{'CRV1':'responseid'})

pf.etable([reg_basic_no_cluster_se, reg_basic, reg_resp_qfe])
```

✓ 0.3s

```
                          est1               est2               est3
------------  -----------------  -----------------  -----------------
depvar                    value              value              value
---------------------------------------------------------------------
Intercept      2.601*** (0.004)   2.601*** (0.010)
treatment     -0.076*** (0.006)  -0.076*** (0.015)  -0.076*** (0.015)
---------------------------------------------------------------------
question                      -                  -                  x
---------------------------------------------------------------------
R2                        0.001              0.001              0.093
S.E. type                hetero     by: responseid     by: responseid
Observations              1E+05              1E+05              1E+05
---------------------------------------------------------------------
Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001
Format of coefficient cell:
Coefficient (Std. Error)
```

Within subject design

- In a 2-cell, within-subject design, participants saw both news items tagged as written by an AI and by a human reporter.
- 1,005 participants for Experiment 2.

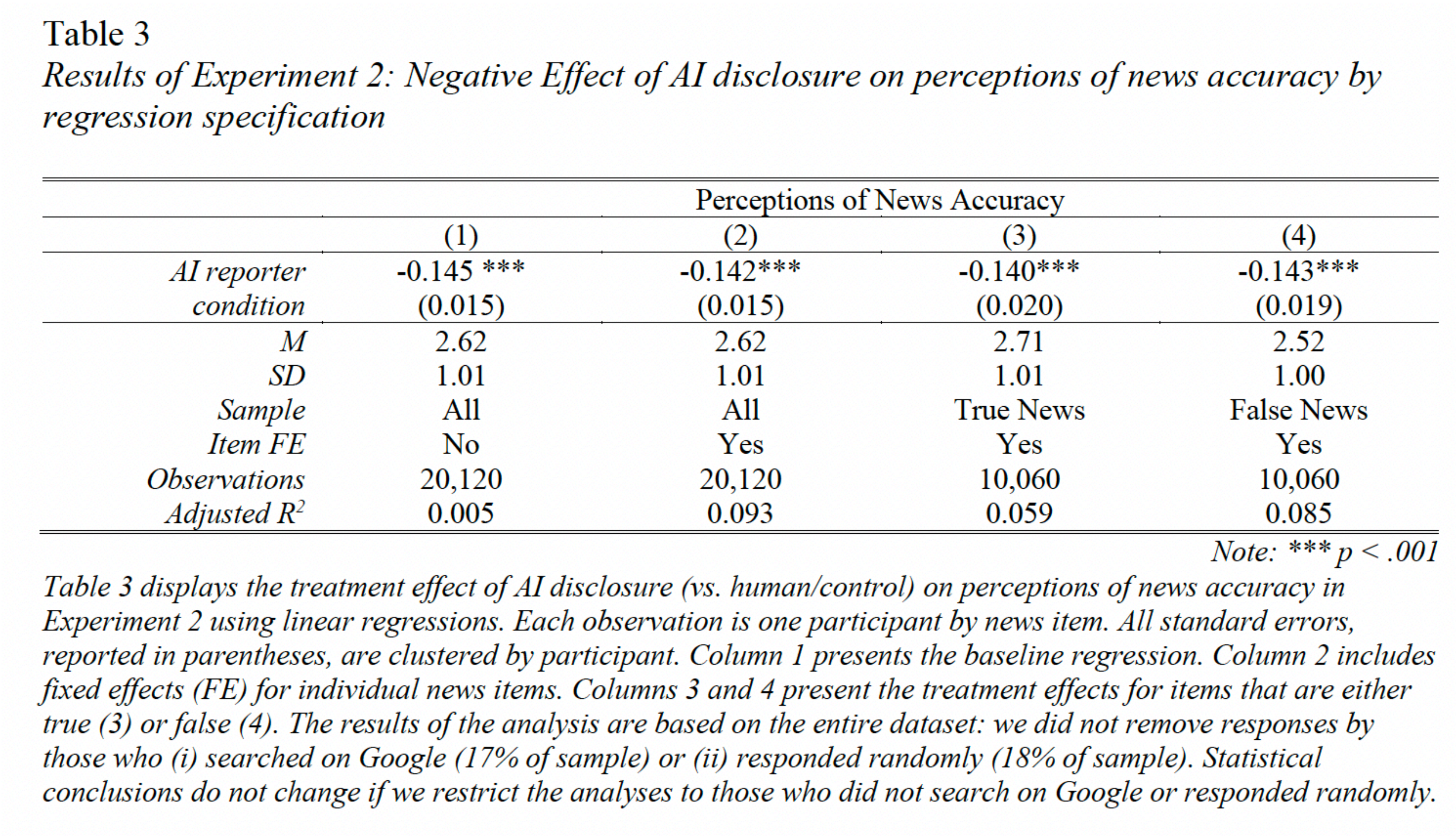


Table 3

Results of Experiment 2: Negative Effect of AI disclosure on perceptions of news accuracy by regression specification

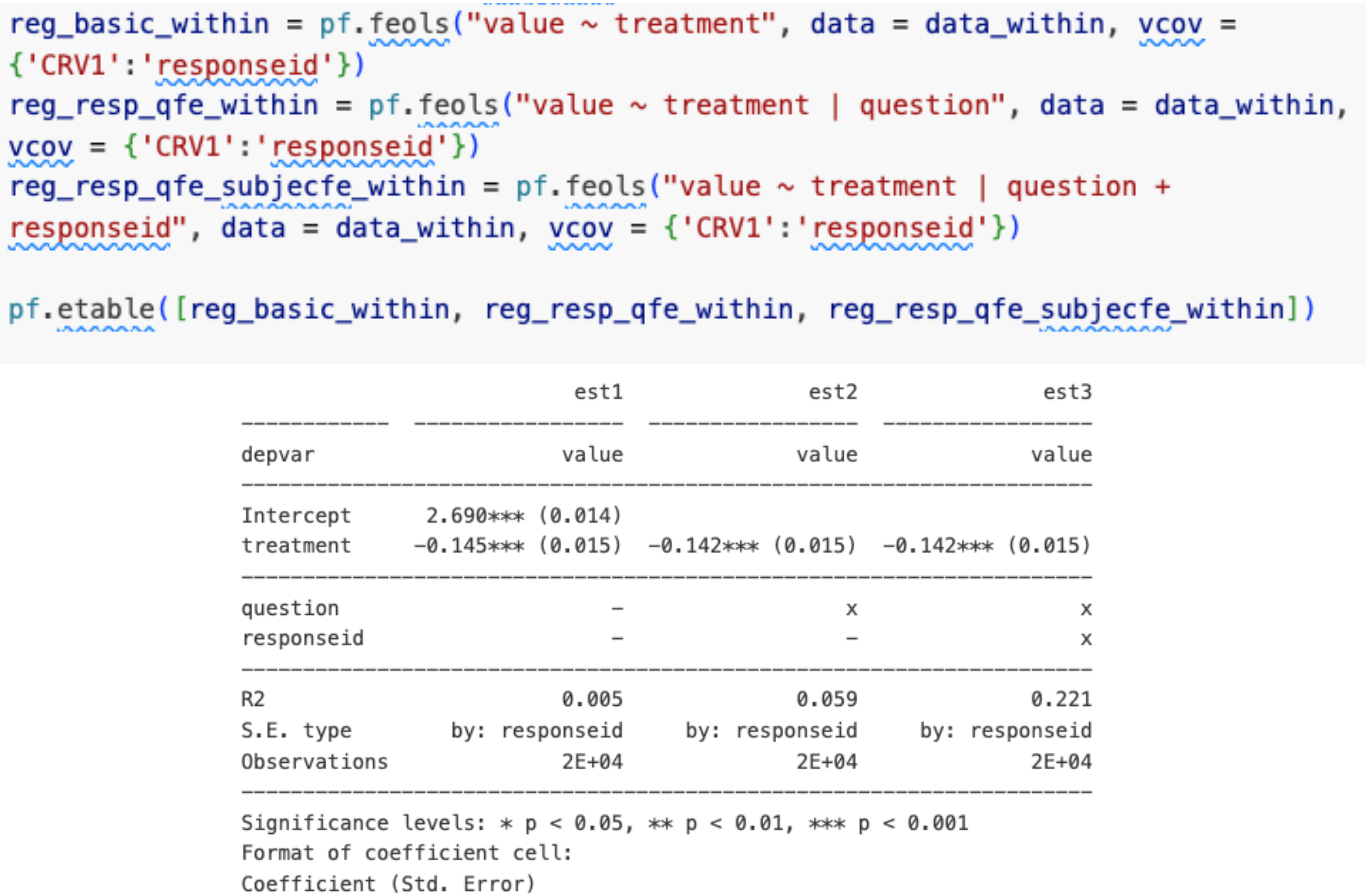
	Perceptions of News Accuracy			
	(1)	(2)	(3)	(4)
AI reporter condition	-0.145 ***	-0.142***	-0.140***	-0.143***
	(0.015)	(0.015)	(0.020)	(0.019)
M	2.62	2.62	2.71	2.52
SD	1.01	1.01	1.01	1.00
Sample	All	All	True News	False News
Item FE	No	Yes	Yes	Yes
Observations	20,120	20,120	10,060	10,060
Adjusted R^2	0.005	0.093	0.059	0.085

*Note: *** $p < .001$*

Table 3 displays the treatment effect of AI disclosure (vs. human/control) on perceptions of news accuracy in Experiment 2 using linear regressions. Each observation is one participant by news item. All standard errors, reported in parentheses, are clustered by participant. Column 1 presents the baseline regression. Column 2 includes fixed effects (FE) for individual news items. Columns 3 and 4 present the treatment effects for items that are either true (3) or false (4). The results of the analysis are based on the entire dataset: we did not remove responses by those who (i) searched on Google (17% of sample) or (ii) responded randomly (18% of sample). Statistical conclusions do not change if we restrict the analyses to those who did not search on Google or responded randomly.

Additional FE in Python

- See ‘course_materials/survey_code/analyze_survey.ipynb’.
-



```
reg_basic_within = pf.feols("value ~ treatment", data = data_within, vcov =
{'CRV1':'responseid'})
reg_resp_qfe_within = pf.feols("value ~ treatment | question", data = data_within,
vcov = {'CRV1':'responseid'})
reg_resp_qfe_subjecfe_within = pf.feols("value ~ treatment | question +
responseid", data = data_within, vcov = {'CRV1':'responseid'})

pf.etable([reg_basic_within, reg_resp_qfe_within, reg_resp_qfe_subjecfe_within])
```

```
                          est1               est2               est3
------------  -----------------  -----------------  -----------------
depvar                   value              value              value
---------------------------------------------------------------------
Intercept      2.690*** (0.014)
treatment     -0.145*** (0.015)  -0.142*** (0.015)  -0.142*** (0.015)
---------------------------------------------------------------------
question                     -                  x                  x
responseid                   -                  -                  x
---------------------------------------------------------------------
R2                       0.005              0.059              0.221
S.E. type       by: responseid     by: responseid     by: responseid
Observations             2E+04              2E+04              2E+04
---------------------------------------------------------------------
Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001
Format of coefficient cell:
Coefficient (Std. Error)
```

Summary

- Survey experiments are useful, even if imperfect.
- Two main types:
  - Across subject.
  - Within subject.
- Example experiment:
  - People view AI generated news headlines as less accurate.